Advanced Regression: 2a Variable selection

Garyfallos Konstantinoudis

Epidemiology and Biostatistics, Imperial College London

7th March 2023

Classical variable or model selection

- Variance explained and ANOVA
- Akaike criterion and other likelihood-based measures

Variable or model selection

Variable or model selection

To select a model (a set of variables, i.e. one or many variables) jointly.

- Focus is not on a single variable but on a model, i.e. one or a combination of many variables.
- Motivation: **To understand** which combination of variables explains best the outcome and **to predict** future outcomes.

Classical variable or model selection

Measures used to compare models:

- Proportion of variance explained
- $F-$statistic and analysis of variances (ANOVA)
- Likelihood based methods
  - Likelihood ratio test (LRT)
  - Akaike information criterion (AIC)
  - Bayesian information criterion (BIC)

Proportion of variance explained

Linear model

$$y = x\beta + \epsilon$$

- Variance decomposition:

$$\underbrace{var(y)}_{\text{Total Variance}} = \underbrace{var(x\beta)}_{\text{Explained Variance}} + \underbrace{var(\epsilon)}_{\text{Error Variance}}$$

- R^2 is the proportion of variance explained by a model

$$R^2 = \frac{var(x\beta)}{var(y)} = 1 - \frac{var(\epsilon)}{Var(y)}$$

Proportion of variance explained

- How to compute? Using sum of squares (SS):
  - Total variance

$$\hat{var}(y) = \frac{1}{n-1} SS_{Total} = \frac{1}{n} \sum_{i=1}^{n} [y_i - \bar{y}]^2$$

  - Explained variance $\hat{y} = x\beta$

$$\hat{var}(\hat{y}) = \frac{1}{n-1} SS_{Explained} = \frac{1}{n} \sum_{i=1}^{n} [\hat{y}_i - \bar{y}]^2$$

  - Error variance

$$\hat{var}(\epsilon) = \frac{1}{n-1} SS_{Error} = \frac{1}{n} \sum_{i=1}^{n} [y_i - \hat{y}_i]^2$$

where the mean is defined as $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

Occam's razor

- When comparing two models it is important not only to consider R^2 but also how complex they are, i.e. how many variables they include.
- Occam's razor (law of parsimony): Simpler solutions are more likely to be correct than complex ones (William of Ockham 1287–1347ad).
- Problem: R^2 will always increase when including more variables.
- Question: Is including more variables actually improving the model fit significantly?

Adjusted proportion of variance explained

- Adjusted R^2

$$R^2_{adj} = 1 - (1 - R) \times \frac{n-1}{n-p-1}$$

- Alternative representation with degrees of freedom (df)

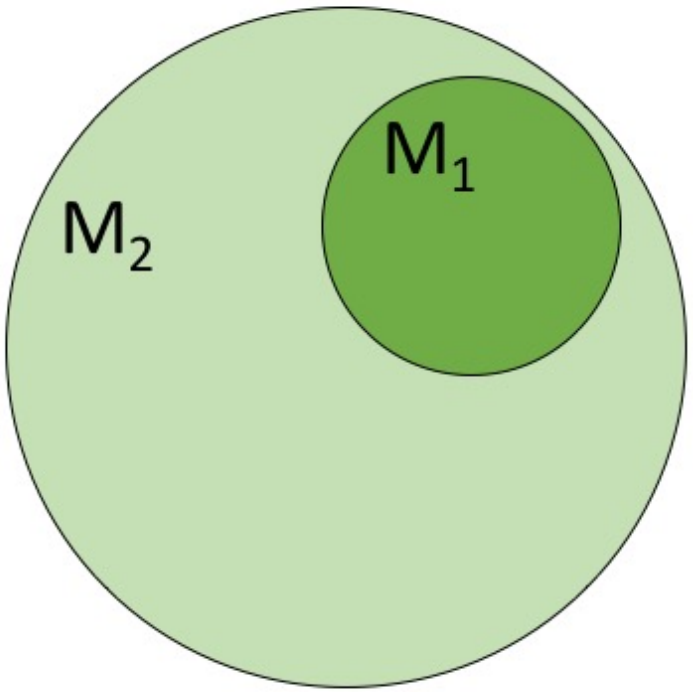
$$R^2_{adj} = 1 - \frac{SS_{Error}/df_e}{SS_{Total}/df_t}$$

where $df_t = n - 1$ and $df_e = n - p - 1$

Analysis of Variances (ANOVA)

What is a nested model?

- Model M_1 is a nested model of model M_2 when model M_2 contains M_1.
- M_1 is a subset of M_2: $M_1 \subset M_2$
- Example: $M_1 = bmi$ and $M_2 = bmi + map$



Analysis of Variances (ANOVA)

- $F-$test, to compare two nested models: a 'full' and a 'reduced' model.
- M_2 'full' model included p_2 predictors.

	df	SS
Regression fit	p_2	$SS_{Explained}(M_2)$
Error	$n - p_2 - 1$	$SS_{Error}(M_2)$
Total	$n - 1$	$SS_{Total}(M_2)$

- M_1 'reduced' model included p_1 predictors, where $p_1 < p_2$.

	df	SS
Regression fit	p_1	$SS_{Explained}(M_1)$
Error	$n - p_1 - 1$	$SS_{Error}(M_1)$
Total	$n - 1$	$SS_{Total}(M_1)$

Analysis of Variances (ANOVA)

- F-test, to compare two nested models:
  - 'full' model (M_2) with p_2 parameter
  - 'reduced' model (M_1) with p_1 parameter
- It will always hold that:
  - $R^2(M1) \leq R^2(M2)$
  - $SS_{Error}(M_1) \geq SS_{Error}(M_2)$
  - $p_2 > p_1$
- But is the 'full' model (M_2) significantly better than a 'reduced' model (M_1)?

Analysis of Variances (ANOVA)

- H_0: Model M_2 fits the data as good as model M_1.

$$F = \frac{(SS_{Error}(M_1) - SS_{Error}(M_2))/(p_2 - p_1)}{SS_{Error}(M_2)/(n - p_2 - 1)}$$

- Under the Null, the test statistic F follows an F-distribution with $(p_2 - p_1)$ and $(n - p_2 - 1)$ degrees of freedom.
- Interpretation 1: If we reject H_0, M_2 fits the data significantly better than model M_1.
- Interpretation 2: By adding more predictors in the complex model compared to the reduced model we can explain more of the variation in Y.
- `anova(M1,M2)` command in `R`.

ANOVA example: Diabetes data

```
> lm1=lm(y~age+sex+map+ltg, data=x)
> lm2=lm(y~age+sex+glu+map+ltg, data=x)
>
> anova(lm1,lm2)
Analysis of Variance Table

Model 1: y ~ age + sex + map + ltg
Model 2: y ~ age + sex + glu + map + ltg
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    437 1610283
2    436 1588468  1     21815 5.9878 0.0148 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- It improves the model fit to add the variable glu to the model.
- The 'full' model (M_2) is better than the 'reduced' model (M_1).

Likelihood ratio test

- The likelihood ratio test (LRT) contrasts two nested models ($M_1 \subset M_2$).
- It is defined as the difference between the log-likelihoods

$$LRT = -2(\log L(M_1) - \log L(M_2))$$

- Note that $-\log L(M_2) \leq \log -L(M_1)$, which is in analogy with $R^2(M_2) \geq R^2(M_1)$.
- The main aim is to test if M_2 provides a significantly better model fit than M_1.

Likelihood ratio test in R

- `library(lmtest)`
- Function: `lrtest(M1,M2)`
- Is there a better model fit when including the predictor glu to the model?

```
> lm1=lm(y~age+sex+map+ltg, data=x)
> lm2=lm(y~age+sex+glu+map+ltg, data=x)
> lrtest(lm1,lm2)
Likelihood ratio test

Model 1: y ~ age + sex + map + ltg
Model 2: y ~ age + sex + glu + map + ltg
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   6 -2439.5
2   7 -2436.5  1 6.0289    0.01407 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Model M_2 (including glu) is marginally better.

Likelihood ratio test in R

- Is there a better model fit when including the predictor map to the model?

```
> lm1=lm(y~age+sex+glu+ltg, data=x)
> lm2=lm(y~age+sex+glu+map+ltg, data=x)
> lrtest(lm1,lm2)
Likelihood ratio test

Model 1: y ~ age + sex + glu + ltg
Model 2: y ~ age + sex + glu + map + ltg
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   6 -2454.4
2   7 -2436.5  1 35.786  2.203e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Model M_2 (including map) is significantly better.
- Model M_2 needs one more parameter to estimate than M_1 (degree of freedom (df)=1).

Akaike information criterion (AIC)

- Akaike information criterion (AIC) combines a measure of model fit with a measure of model complexity.

$$AIC = -2 \log L + 2p$$

  - L Maximum likelihood of the model
  - p Model complexity: Number of parameters in the model
- The best model is the one with the minimum AIC value (minimum information loss).
- The AIC can be used for model comparison, not to assess the quality of the model fit.
- `AIC(M1,M2)` command in R.

Bayesian information criterion (BIC)

- Also the Bayes information criterion (BIC) combines a measure of model fit with a measure of model complexity.

$$BIC = -2 \log L + \log(n)p$$

  - L Maximum likelihood of the model
  - p Model complexity: Number of parameters in the model
- The best model is the one with the minimum BIC value (minimum information loss).
- The BIC can be used for model comparison, not to assess the quality of the model fit.
- `BIC(M1,M2)` command in R.

AIC and BIC

- More generally, we can understand information criteria (IC) as a compromise between model fit and model complexity

$$IC = -2logL + k \times p$$

  - L Model fit: Maximum likelihood of the model
  - p Model complexity: Number of parameters in the model
- The best model is the one with the minimum IC value (minimum information loss).
- k defines the penalty of the model complexity
  - AIC: $k = 2$
  - BIC: $k = \log(n)$

AIC and BIC in R

```
> lm1=lm(y~age+sex+map+ltg, data=x)
> lm2=lm(y~age+sex+glu+map+ltg, data=x)
>
>
> AIC(lm1,lm2)
    df      AIC
lm1  6 4891.012
lm2  7 4886.983
>
> BIC(lm1,lm2)
    df      BIC
lm1  6 4915.560
lm2  7 4915.622
```

- There is no consensus between AIC and BIC.
- Using the AIC we would prefer M_2, but using the BIC we would prefer M_1.
- This is not a strong evidence that adding the variable glu (glucose) has a lot of benefit.

AIC and BIC in R

```
> lm1=lm(y~age+sex+glu+ltg, data=x)
> lm2=lm(y~age+sex+glu+map+ltg, data=x)
>
>
> AIC(lm1,lm2)
    df      AIC
lm1  6 4920.769
lm2  7 4886.983
>
> BIC(lm1,lm2)
    df      BIC
lm1  6 4945.316
lm2  7 4915.622
```

- In contrast the variable map (blood pressure) greatly improves the model fit.
- Since the variable map (blood pressure) is supported by both methods we have greater confidence that it improves the model fit.

How to decide which models to test?

- Backward selection
  1. Start with the full model and include all p variables available.
  2. Identify the variable with the weakest evidence and remove it.
  3. Evaluate the model with $p - 1$ variables.
  4. Identify the variable with the weakest evidence and remove it.
  5. ...
- Forward selection
  1. Start to evaluate all models including just a single predictor variable.
  2. Identify the variable with the strongest univariable impact.
  3. Evaluate all models including the best single predictor variable and one additional variable.
  4. Identify the tuple of two variables with the strongest impact.
  5. ...

Alternatives for model selection

Warning!

Backward and forward selection do rarely agree. They are highly instable and there is no guaranty that they find the optimal model. It is not recommended to use them.

Alternatives for model selection:

- Evaluate all possible models: Becomes computationally infeasible even with a moderate number of variables.
  - $p = 10$ variables have $2^{10} = 1,024$ possible models
  - $p = 20$ variables have $2^{20} = 1,048,576$ possible models
- Penalised regression (Lecture 3b)